Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

Impact of Scanning and Optical Character Recognition on the Digital Transformation of Research Publications in Bayero University Library, Kano, Nigeria

[1,2,*]Moruf Hawwau A., [2]Enang Uduak U., [2]Umoren Eboro E.
[1]Department of Library and Information Science, Federal University Dutsin-Ma, Nigeria
[2]Department of Library and Information Science, University of Uyo, Nigeria
hawwau.moruf@yahoo.com

Abstract
Scanning is a fundamental step in the digitization workflow, allowing physical documents to be transformed into a format that can be stored, manipulated, and accessed electronically. This study was carried out to investigate the impact of document scanning and Optical Character Recognition (OCR) on the visibility of research publications at Bayero University, Kano. The study was guided by four specific objectives with their matching research questions. The design employed in this study was survey research design. Data were collected using structured questionnaire with Four point Likert Scale. The data were analyzed using descriptive statistics of mean and standard deviation, while Student-t-test was used to test the hypotheses at 0.05 level of significance. The results revealed that the mean range for document scanning was 1.86 - 1.99, with a standard deviation range of 1.29 - 1.34, indicating close clustering around the weighted mean. Notably, all mean values were below 2.0, suggesting a relatively limited impact. The use of scanning software and Photoshop Elements for batch processing also showed mean and standard deviation values of 1.97 and 1.34, contributing to a pooled mean of 1.93, indicating a lesser extent of impact. For OCR, the mean range was 1.63 - 1.95, with a standard deviation range of 1.17 - 1.38. The mean responses were closely grouped below 2.0, and the pooled mean is 1.85, indicating consensus among respondents that OCR has a positive impact on visibility, albeit to a lesser extent. Statistical analyses support these findings, with the t-test for document scanning yielding a t-value of 25.71 and a probability value of .0001 at the .05 alpha level, indicating statistical significance. The null hypothesis was rejected, confirming a significant impact of document scanning on visibility. Similarly, the t-test for OCR produces a t-value of 26.93, with a probability value of .001, reinforcing the statistical significance and rejecting the null hypothesis. In conclusion, both document scanning and OCR contributes to the visibility of research publications at Bayero University, Kano, the impact is observed to be moderate. The paper recommends that institutions should take advantage of cloud services (Softwae As A Service, SAAS) for document scanning, as there are off-the shelf and cloud services that can help institutions scan and prepare documents.

Keywords: Academic library, Digitization Workflow, Scanning, Digital transformation Research Publication, University.

Introduction

Scanning is the electronic conversion of a physical document or book into a digital image. In digitization process, document scanning refer to the process of converting paper documents into electronic documents by capturing valuable information, and saving the document in a central repository for easy retrieval later (Safonov *et al.*, 2019). The result of scanning is typically a series of image files, such as JPEG or TIFF, which faithfully replicate the visual appearance of the original document. Scanning is useful for preserving the visual layout, graphics, and other elements of a document, but the resulting files are essentially images and lack the ability to recognize or interpret the text. This process not only helps in saving space and costs, but also enhances accessibility, security, and regulatory compliance (Azim *et al.*,

Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

2018). Document scanning plays a crucial role in digitization process, because it reduces storage requirements, improves accessibility to information, and helps protect data by creating a digital backup of physical documents. However, simply scanning documents alone is not enough to be called digitization. That is because this method only creates a single searchable "index", (the file name), which can make it more difficult to retrieve any given file.

By understanding the concepts of scanning and digitization, as well as activities involved in each, libraries can make informed decisions and develop an effective document digitization management strategy that aligns with their specific goals. According to Lapworth and Chung (2021), scanning involves using a scanning device, such as a scanner or camera, which captures the content of the document by creating a digital image of each page. The scanner uses light sensors to detect and record the variations in light and dark on the document, producing a visual representation of the text and images. The scanned image typically preserves the visual characteristics of the original, replicating text, graphics, or other visual elements. Scanning is a fundamental step in the digitization workflow, allowing physical documents to be transformed into a format that can be stored, manipulated, and accessed electronically.

While both document scanning and digitization transform physical documents into digital files, the end result differs in functionality and output. Document scanning helps creates a high quality image of a document though the resulting files, which are not editable or searchable; it is perfect for archiving historical documents or creating digital one-to-one copies of physical documents (Terras, 2015). Digitization on the other hand, adds searchable metadata to files to make them searchable by text; it often involves Optical Character Recognition (OCR) Technology to digitize the contents of the document and to create editable files that is ideal for frequent access (Christy *et al*., 2017). According to Mizumoto and Yilmaz (2018), digitization is the process of converting printed and handwritten text contained on a scanned document into a digital format that is readable by computers using OCR technology. Scanning being the important stage in digitization workflow starts with creating a digital image of the original document. Then, OCR Software is used to convert the contents of the document into machine-encoded text. This allows for keyword searching, text editing, and data mining, and taking data manipulation and digitization to a completely new level.

Optical Character Recognition is a technology that goes beyond scanning by adding the ability to recognize and convert text within images into machine-readable text. The primary objective of OCR is to recognize and extract the text information from the visual representations, the turning it into editable and searchable data (Abdelaziz and Fazil, 2023). This technology is widely used to digitize printed documents, making the content accessible for electronic editing, indexing, and search functionalities. OCR plays a crucial role in automating data entry processes, improving the efficiency of document management, and facilitating the integration of paper-based information into digital workflows (Reul, 2020). Ruhaimi *et al*. (2018) maintained that OCR software thus helps to analyse the shapes, patterns, and arrangements of characters in a document, whether it's a printed page, a scanned image, or a photograph containing text. OCR software analyzes the scanned images, identifies patterns that correspond to letters and words, and translates them into editable and searchable text. The key difference with OCR is that it transforms the scanned images into text-based documents (e.g., PDF, Word, or plain text files), allowing users to search, edit, and manipulate the content (Ruhaimi *et al.*, 2018). OCR is crucial for making digitized documents more accessible and functional, as it enables text extraction and manipulation. In summary, scanning is the process of creating digital images of physical documents, while OCR is a subsequent step that involves converting those images into machine-readable text, adding a layer of functionality and search ability to the digitized content. (de Oliveira, 2023). Implementing a document scanning and OCR in

Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

digitization strategy offers numerous benefits to libraries that include: saving space; improving accessibility: improve security; cost saving; increasing efficiency; regulatory compliance (Nguyen, 2020). Their impacts on digitization processes is multifaceted, bringing about significant advancements in the way information is transformed from physical to digital formats (Rao, 2016).

Research is carried out in the university in other to facilitate learning and advance knowledge. Libraries in public universities have become the repertoire for research publications management through their institutional repositories. Most of the researches submitted to the university libraries nowadays are produced in hard copies where the soft copies are hardly available for use. It is also observed that most researchers prefer submitting published researches to their institutions. This thereby required the university libraries to convert and make them available through digitization. Scanning constitutes an important stage in digitization workflow of any university library. On this basis, this study examined the impact of document scanning on scholarly research publications in Bayero University libraries, Kano.

Background of the Study

Bayero University, Kano (BUK), established in 1975 as one of Nigeria's pioneer universities, stands as a prominent non-profit public institution in Kano, Northwest Nigeria. With 17 faculties offering diverse academic disciplines, BUK plays a pivotal role in providing both single and double honours degrees to its students. The information presented is based on the Bayero University, Kano Annual Report of 2019. Central to its academic mission, the university library serves as a hub for teaching, learning, and research activities within the university community. The library's robust collection of Digital Information Resources (DIRs) includes e-books, e-journals, e-databases, digitized theses, e-reference materials, e-newspapers, digitized past question papers, CD-ROM databases, digitized manuscripts, and Compact Discs. Access to these resources is facilitated with the library's ICT facilities on-site or via personal computers, smartphones, and laptops. Authorized users are granted access through a secure system involving unique usernames and passwords issued by the university library's Automation and Multimedia Department.

Recognizing the importance of guiding users, the library conducts orientation and user education sessions for all students. Additionally, awareness of the available digital resources is disseminated through the university bulletin, libri updates, and informative pamphlets that serve as a guide to utilizing the library effectively. To monitor resource usage, the university library diligently maintains a usage statistics register for both traditional print materials and digital resources. This practice aids in tracking and evaluating the utilization of the library's vast resources, ensuring that the needs of the academic community are met efficiently (Abdullahi, 2020). Overall, BUK's commitment to providing diverse and accessible resources underscores its dedication to fostering a conducive environment for academic excellence.

Purpose of the Study

The general objective of the study is to evaluate the impact of scanning and optical character recognition on the digital transformation of research publications in Bayero University Library, Kano. The specific objectives are:

I. To determine the effect of document scanning on the visibility of academic research publications in Bayero University Library, Kano.
II. To examine the use of Optical Character Recognition, affect visibility of research publications in Bayero University Library, Kano.
III. To investigate the relationship between document scanning in digitization and the visibility of academic research publications in Bayero University Library, Kano

Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

IV. To investigate the relationship between Optical Character Recognition and the visibility of academic research publications in Bayero University Library, Kano.

Research Questions

This study was guided by the following research questions:

1. What is the effect of document scanning in digitization on the visibility of research publications in Bayero University Library, Kano?
2. What is the effect of Optical Character Recognition on the visibility of research publications in Bayero University Library, Kano?
3. What is the relationship between document scanning in digitization and the visibility of research publications in Bayero University, Kano?
4. What is the relationship between Optical Character Recognition and the visibility of research publications in Bayero University, Kano?

Research Hypotheses

The following null hypotheses were tested at 0.05 level of significance.

1. There is no significant relationship between document scanning in digitization and the visibility of research publications in Bayero University, Kano
2. There is no significant relationship between Optical Character Recognition and the visibility of academic research publications in Bayero University, Kano.

Literature Review

This review was done on scanning and Optical Character Recognition as constructs of digitization workflow. It further focused on the influence of these constructs on the visibility of research publications.

In a related work, Courtney (2020) studied the use of CleanPage, a novel smartphone- based document and whiteboard scanning system, for the high-quality digitization of content from pages and whiteboards to sharable online material. Courtney found that CleanPage requires one button-tap to capture, identify, crop, and clean an image of a page or whiteboard. Unlike equivalent systems (other scanning systems), no user intervention is required during processing, and the result showed a high-contrast, low-noise image with a clean homogenous background. The similarity is that the prior study used comparative analysis in establishing the effect of scanning on user experience, while the present study looks at the influence of scanning on visibility.

Akinbade *et al*. (2020) conducted a research on using an adaptive thresholding algorithm-based Optical Character Recognition system for information extraction from images with complex backgrounds. The results showed that the system was able to extract English character-based texts from images with complex backgrounds with 69.7% word-level accuracy and 81.9% character- level accuracy. The proposed method in this study proved to be more efficient as it outperformed the existing methods in terms of the character level percentage accuracy. The study concluded that algorithm-based Optical Character Recognition system can be recommended for digitizing scholarly documents with complex background images for quality text extraction in the course of scanning and reproducing scholarly publications. The two studies relate in dealing with the same independent variable which is scanning. However, they are at variance in relating scanning with dependent variable as the prior study concentrated on comparing the veracity of scanning through algorithm-based optical character recognition, while the current study measures the influence of scanning on visibility.

Jain *et al.* (2021) the research provides a comparative study of various OCR toolsets among the proprietary, open-source and online OCR software. The result of evaluation revealed that

Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

different OCR tools have different capabilities, and a single OCR toolset may not fit in all the domains. Since image quality plays an important role in text recognition, the study concluded that the need to carefully choose the use of OCR tools in digitization of scholarly publications to enhance their visibility is important. Both studies examined scanning as a variable however, they employed different methodology.

Arief *et al*. (2018) investigated the use of Google Vision OCR in Apache Hadoop Environment for automated extraction of large scale scanned document images. It was found that automated extraction systems can recognize text in a large-scale image document accurately and can be operated in a real-time environment. The study is experimental and established the connection between quality scanning and visibility through automated extraction system for extracting text from a large-scale scanned document images using modern OCR technology. The two studies relate in terms of variables treated but differ in their research objectives and methodologies.

Zhou (2023) carried out a review on document image enhancement based on document degradation problem. The study found that when facing the problem of fading, a model for stroke connectivity can be used, while the other three degradation problems are mostly deep learning models. Both studies are related in their aims to optimising scanning process but differ in their research approach, design and methodology.

Methodology

The design employed in this study was survey research design. The purpose of survey design is to describe the current condition of an area of study. Rubin and Babbie (2016) stated that survey design is a research methodology used to gather information about prevalence, distribution, or patterns of certain variables or phenomena within a specific population. Questionnaire, as instrument was used to elicit information on the dependent and independent variables as well as answer the research question and test the hypotheses. Four point Likert scale was employed in order to understand respondents' opinion in respect to the extent of impact without being neutral on a specific item. The questionnaire was structured on four response scale ratio of Less Extent (LE), Moderate Extent (ME), Great Extent (GE), Very Great Extent (VGE), and the scales attract the following points: LE-1 point; ME-2 points, GE-3 points while VGE- 4 points. A total enumeration sampling method was employed, since the population size was manageable. A total of 106 respondents in Bayero University Library, Kano completed and returned the questionnaire administered to them. Descriptive statistics was used to answer the research questions. In deciding the mean and answers to the research questions, the weighted options were:

Very Great Extent (VGE)	3.50-4.0
Great Extent (GE)	3.00-3.49
Moderate Extent (ME)	2.50-2.99
Less Extent (LE)	1.00-2.49

Student-t-test was used to test the hypotheses at 0.05 level of significance.

Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

Results

Research Question 1: What is the effect of document scanning in digitization on the visibility of research publications in Bayero University Library, Kano?

Table 1: Mean and Standard Deviation of Respondents on the Impact of Document Scanning on the Digital Transformation of Research Publications

S/N	Extent to which scanning techniques impact visibility of research publications	Mean	Std. Dev	Remarks
1	Converting data from non-machine readable to machine readable with scanner	1.94	1.30	Less Extent
2	Creating digital surrogates	1.91	1.34	Less Extent
3	Setting appropriate file size for image display	1.99	1.36	Less Extent
4	Creating tagged image files formats (TIFF)	1.93	1.30	Less Extent
5	Using JPEG 2000 (Joint Photographic Experts Group)	1.93	1.29	Less Extent
6	Balancing file size and image quality	1.93	1.33	Less Extent
7	Using scanning software for editing	1.97	1.34	Less Extent
8	Using Photoshop elements (software) for batch processing	1.99	1.34	Less Extent
9	scanning can equip publications with accessibility features like text-to-speech and adjustable font sizes	1.86	1.32	Less Extent
10	High-quality scans ensure clear and legible text and images of publications	1.88	1.32	Less Extent
	Pooled Mean and Std. Dev.	1.93	0.02	Less Extent

Table 1 gives the summary of the mean and item analysis of respondents on the impact of document scanning techniques on visibility of research publications. The result shows that the mean range is 1.86 -1.99. The standard deviation range is 1.29 -1.34. This indicates that the means are close to each other and cluster around the weighted mean. The result also shows that all the mean values are below 2.0, indicating less extent of impact. The results also discovered that use of scanning software was to a less extent which showed a mean and standard deviation of 1.97 and 1.34 respectively. Using Photoshop elements (software) for batch processing also revealed to a less extent. The pooled mean value is 1.93. This reveals that there is less impact of document scanning on visibility research publications.

Research Question 2: What is the effect of Optical Character Recognition on the visibility of research publications in Bayero University Library, Kano?

Table 2: Mean of respondents on impact of Optical Character Recognition on the visibility of research publications.

S/N	Extent to which the following OCR attributes impact visibility of research publications	Mean	Std. Dev	Remarks
1	Optical character Recognition (OCR) enables the conversion of scanned research publication of text into machine-readable and searchable text	1.79	1.38	Less Extent
2	OCR allows for the extraction of text from scanned research publication, making the content accessible in a format that users can manipulate.	1.93	1.33	Less Extent
3	OCR facilitates the integration of text-to-speech technology of research publications.	1.82	1.27	Less Extent
4	OCR transforms scanned images into editable text, enabling users to make annotations, corrections, or modifications to the research publication.	1.95	1.26	Less Extent

Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

5	With OCR, the digitized research publication can be easily translated into different languages using automated translation services.	1.88	1.32	Less Extent
6	OCR-generated text allows for data mining and analysis.	1.63	1.28	Less Extent
7	Using OCR system for extractions from research publication with complex backgrounds.	1.91	1.28	Less Extent
8	OCR contributes to the creation of navigational aids such as hyperlinks, bookmarks, and a table of contents	1.88	1.31	Less Extent
9	While OCR primarily focuses on extracting and recognizing text, efforts are made to preserve the original layout and formatting of research publication.	1.87	1.29	Less Extent
10	Optical character Recognition (OCR) software to extract text from images is attached to scanner for quality extraction	1.86	1.17	Less Extent
	Pooled Mean and Std. Dev.	1.85	0.05	Less Extent

Table 2 gives the summary of the mean and item analysis of respondents on impact of Optical Character Recognition on the visibility of research publications. The result shows that the mean range is 1.63 -1.95. The standard deviation range is 1.17 -1.38. This indicates that the mean responses are close to each other. The result also shows that all the items have mean responses below 2.0. The pooled mean is 1.85. This indicates that all the respondents agreed that to a less extent, OCR has a positive impact on visibility of research publications.

Research Question 3: What is the relationship between document scanning in digitization and the visibility of research publications in Bayero University Library, Kano?
Ho1 There is no significant impact of scanning on visibility of research publications in Bayero University, Kano.

Table 3: Summary of t-test analysis for significant impact of document scanning on visibility of research publications in Bayero University, Kano

Variables	Mean	Std. Dev	T	*P*-value	Decision
Document scanning	1.93	0.02	25.71	.0001	*
visibility of research publications	3.58	0.78			

*Significant at $p<.05_{.0001}$

Table 3 presents the summary of the t-test analysis. The result shows that the t-value is 25.71. The probability value at .05 alpha level is .0001. Since the *P*-value is less than the alpha level, the result is statistically significant. The null hypothesis is rejected, thus, there is a significant impact of document scanning on visibility of research publications in Bayero University, Kano.

Research Question 4: What is the relationship between Optical Character Recognition and the visibility of research publications in Bayero University, Kano?

Ho2 There is no significant influence of Optical Character Recognition on visibility of research publications in Bayero University, Kano.

Jewel Journal of Librarianship
ISSN: 2141-3908 (Print); ISSN: 2736-0881 (Online)
Volume 19, Issue 1; Published: March, 2024
https://www.jeweljournals.com

Table 4. Summary of t-test analysis for significant impact of Optical Character Recognition on visibility of research publications in Bayero University, Kano

Variables	Mean	Std. Dev	T	*P*-value	Decision
Document metadata and indexing	1.85	0.05	26.93	.001	*
visibility of scholarly research publications	3.58	0.78			

*Significant at $p<.05_{.001}$

Table 4. presents the summary of the t-test analysis. The result shows that the t-value is 26.93. The probability value at .05 alpha level is .001. Since the *P*-value is less than the alpha level, the result is statistically significant. Therefore, the null hypothesis rejected. Thus, there is a significant impact of Optical Character Recognition on visibility of research publications in Bayero University, Kano.

Discussion of Findings

The findings of this study showed all the mean values to be below 2.0, indicating less extent of impact. The range of the standard deviation indicates that the means are close to each other and cluster around the weighted mean. The pooled mean value is also below 2.0. This reveals that there is less impact of document scanning on visibility of research publications Bayero University, Kano. The hypothesis test however, shows that there is a significant impact of document scanning on visibility of research publications in Bayero University, Kano. The technology that supports digital scanning might just be finding its way into Nigeria. However, even though its scanning penetration is low, it still has a significant impact on document visibility. This finding is corroborated by Courtney (2020) who studied the use of Clean Page, a novel smartphone-based document and whiteboard scanning system, for the high-quality digitization of content from pages and whiteboards to sharable online material. Comparative results were presented for a selection of scenarios showing the versatility of the design. The user experience for each scanning app was assessed, showing Clean Page to be fast and easier to use. The study revealed that resultant effect of scanning on user experience when scanned document is visible to users.

The finding on the impact of Optical Character Recognition on the visibility of research publications showed standard deviation range indicating that the mean responses are close to each other. The result also shows that all the items have mean responses below 2.0. The pooled mean is 1.85. This indicates that all the respondents agreed that to a less extent, OCR attributes have a positive impact on visibility of research publications. Further test of hypothesis shows that there is a significant impact of OCR on visibility of research publications in in Bayero University, Kano, Nigeria. This finding is supported by Jain *et al.* (2021) who did a comparative study of various OCR toolsets among the proprietary, open-source and online OCR software. A variety of parameters was employed for comparative analysis including; suitability of operating system, multilingual support, input formats supported and publications formats supported. The result of evaluation revealed that different OCR tools have different capabilities, and a single OCR toolset may not fit in all the domains. Since image quality plays an important role in text recognition, the study concluded that the need to carefully choose the use of OCR tools in digitization of scholarly publications to enhance their visibility is important. This finding is also supported by Arief *et al*. (2018) which revealed that Google Vision OCR shows better extraction performance compared to other OCR tools. It was found that automated extraction systems can recognize text in a large-scale image document accurately and can be operated in a real-time environment. The study is experimental and established the connection between quality scanning and visibility through automated

extraction system for extracting text from a large-scale scanned document images using modern OCR technology to improve visibility and discoverability of digitized documents. In the same vein, Zhou *et al.* (2023) found that banalization model can be used to deal with the degradation of background textures, lettering smudges in enhancing scanning. The study found that when facing the problem of fading, a model for stroke connectivity can be used, while the other three degradation problems are mostly deep learning models. These methods can handle specific types of degradation problems very well based on machine learning methods in scanning process.

Conclusion and Recommendations

Based on the findings of the study, it can be concluded that digitization workflow is an on-going process in universities and it is happening at a moderate rate. The outcome of the study shows that there is a moderate impact of both document scanning and OCR on visibility of research publications in Bayero University, Kano Nigeria.

Based on the findings of the study, the research recommends that:

1. Tertiary institutions should have clear goals and objectives as regards digitization of information and library services
2. There are off-the shelf and cloud services that can help institutions scan and prepare documents. Institutions should take advantage of cloud services (Softwae As A Service, SAAS) for document preparation and scanning.
3. Universities should also provide repositories, where articles and publications from staff and students are hosted. This will not only enhance the visibility of publications, but also the standing of the university as well

References

Abdelaziz, T. A. I. and Fazil, U. (2023). Applications of integration of AI-based Optical Character Recognition (OCR) and Generative AI in Document Understanding and Processing. *Applied Research in Artificial Intelligence and Cloud Computing*, 6(11): 1-16.

Abdullahi, L. A. (2020). Assessment of Digital Literacy Competencies of Undergraduate students of Bayero University Kano, Kano State, Nigeria. *Jewel Journal of Librarianship*, 15(1):76-89.

Akinbade, D., Ogunde, A. O., Odim, M. O. and Oguntunde, B. O. (2020). An adaptive thresholding algorithm-based optical character recognition system for information extraction in complex images. *Journal of Computer Science*, 16(6): 784-801.

Arief, R., Mutiara, A.B., Kusuma, T.M. and Hustinawaty (2018). Automated Extraction of Large Scale Scanned Document Images using Google Vision OCR in Apache Hadoop Environment. *International Journal of Advanced Computer Science and Applications*, 11(9): 112-116.

Azim, N., Mat Yatin, S. F., Jensonray, R. and Ayub Mansor, S. (2018). Digitization of records and archives: Issues and Concerns. *International Journal of Academic Research in Business and social sciences*, 8(9): 170-178.

Christy, M., Gupta, A., Grumbach, E., Mandell, L., Furuta, R. and Gutierrez-Osuna, R. (2017). Mass digitization of early modern texts with optical character recognition. Journal on Computing and Cultural Heritage (JOCCH), 11(1): 1-25.

Courtney, J. (2020). CleanPage: Fast and clean document and whiteboard capture. *Journal of Imaging*, 6: 1-17. doi:10.3390/jimaging6100102, www.mdpi.com/journal/jimaging.

de Oliveira, L. L., Vargas, D. S., Alexandre, A. M. A., Cordeiro, F. C., Gomes, D. D. S. M., Rodrigues, M. D. C. and Moreira, V. P. (2023). Evaluating and mitigating the impact of OCR errors on information retrieval. *International Journal on Digital Libraries*, 24(1), 45-62.

Jain, P., Taneja, K. and Taneja, H. (2021). Which OCR toolset is good and why: A comparative study. *Kuwait Journal of Science*, 48(2):1-9.

Lapworth, E. and Chung, S. K. (2021). The Archives at the Tip of Their Fingers: Exploring User Reactions to Large-Scale Digitization. *Journal of Archival Organization*, 18(1-2): 1-36.

Mizumoto, R. M. and Yilmaz, B. (2018). Intraoral scan bodies in implant dentistry: A systematic review. *The Journal of prosthetic dentistry*, 120(3): 343-352.

Nguyen, N. K., Boros, E., Lejeune, G. and Doucet, A. (2020). Impact analysis of document digitization on event extraction. In 4th workshop on natural language for artificial intelligence (NL4AI, November 2020) co-located with the 19th International conference of the Italian Association for artificial intelligence (AI* IA 2020) (Vol. 2735, pp. 17-28).

Rao, N. V., Sastry, A. S. C. S., Chakravarthy, A. S. N. and Kalyanchakravarthi, P. (2016). Optical Character Recognition Technique Algorithms. *Journal of Theoretical & Applied Information Technology*, 83(2):12-21.

Reul, C. (2020). An intelligent semi-automatic workflow for Optical Character Recognition of Historical Printings. *Bayerische Julius-Maximilians-Universitaet Wuerzburg* (Germany).

Rubin, A. and Babbie, E. R. (2016). Empowerment series: Research methods for social work. Cengage Learning.

Ruhaimi, N. N. A., Yatin, S. F. M. and Fadzil, N. H. M. (2018). Scanning Process in Digitization of Records and Archives Materials. *International Journal of Academic Research in Business and Social Sciences*, 8(9): 191–201.

Safonov, I. V., Kurilin, I. V., Rychagov, M. N. and Tolstaya, E. V. (2019). *Document Image Processing for Scanning and Printing*. Springer International Publishing.

Terras, M. (2015). Cultural heritage information: Artefacts and digitization technologies. *Cultural heritage information: Access and management*, 63-88.

Zhou, Y., Zuo, S., Yang, Z., He, J., Shi, J. and Zhang, R. (2023). A Review of document image enhancement based on document degradation problem. *Applied Sciences*, 13(13): 7855.